bpifrance
SERVIR L'AVENIR

ESTIMATION DE L'EMPREINTE CARBONE AU TRAVERS DU "MACHINE LEARNING" ET DES TECHNIQUES DES REGRESSIONS

INTRODUCTION

- **Ce travail s'inscrit dans le cadre de l'estimation de l'empreinte carbone des entreprises du portefeuille de BPI France.**

- Il vise à prédire les émissions scope 1 et 2 2023 au travers du « machine learning » et des techniques de régressions

- **Analyses séquencées trois parties :**
  - Première partie centrée sur la mise en contexte de l'approche par la prédiction. Elle inclura :
    - *Quels sont les objectifs derrière un approche par la prédiction de l'empreinte carbone au travers le machine learning et les régressions?*
    - *La définition du « machine learning » et ses principes*
    - Une brève présentation de la méthodologie du calcul de l'empreinte
  - Deuxième partie sera centré sur la présentation des différents modèles de régression

Nous aborderons la régression linéaire, les régressions pénalisées, ainsi que la métrique de performance utilisée. La présentation de la capacité prédictive de chaque modèle

  - La troisième partie se concentrera sur les limites et les extensions de ce travail

MISE EN CONTEXTE

❑Problématiques

- Recherche de données sur les émissions carbone des entreprises qui ne déclarent pas.
- Quantification difficile en présence de données incomplètes ou absentes.
- Besoin de quantifier l'empreinte attribuable du portefeuille de Bpifrance
- Limites dans les méthodes traditionnelles (imputation par la médiane)

❑Objectifs clés

- Estimer les émissions scope 1 et 2 avec des techniques statistiques
- Prédire les émissions scope 1 &2 sur l'année 2023.
- Tester la réplicabilité des résultats sur de nouvelles données avec le machine learning

MISE EN CONTEXTE

❑ Définition et principes du ML dans le cadre de l'apprentissage supervisé

- Prédire une réponse à partir de données d'entrées
- Formé sur un ensemble de données appelées données d'apprentissages (90% de la base de données)
- Prédiction sur un ensemble de test (10% de la base de données)
- Évaluation de la réplicabilité sur de nouvelles données

PRÉSENTATION DE REUTERS

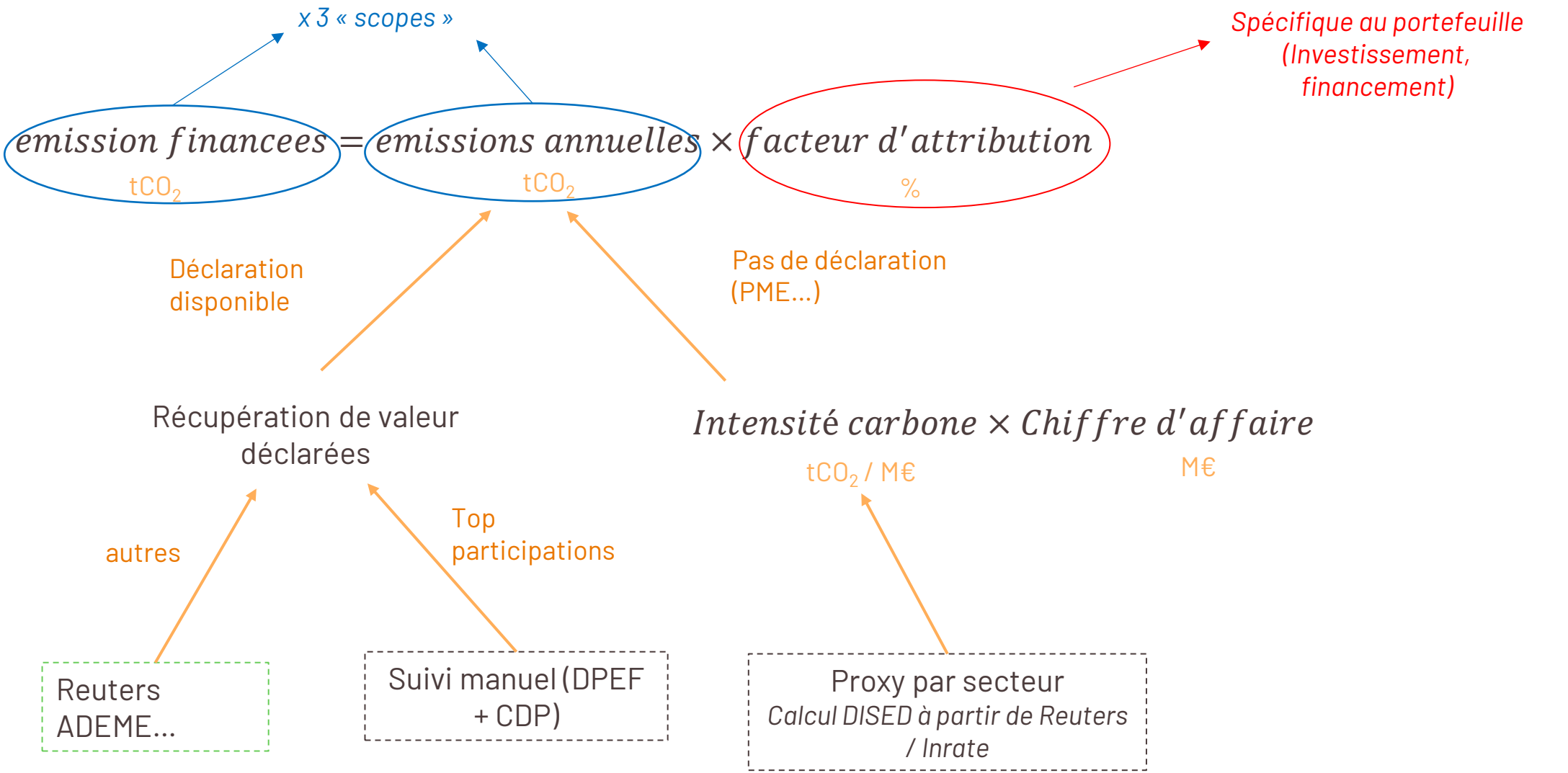
- Fournisseur mondial américano-britannique des données sur les marchés financiers.
- Données sur les entreprises internationales selon la classification naice, trbc.
- La base de données contient des informations sur le nom des entreprises, les classifications sectorielles, le chiffre d'affaires, les émission scopes 1,2, 3, sur la région etc.

	name_or_code	revenue_usd	ric	isin	year	country
1	Rebosis Property Fund Ltd	91224862	REBJ.J	ZAE000201687	2023	south africa
2	Rebosis Property Fund Ltd	91224862	REBJ.J	ZAE000201687	2022	south africa
3	Rebosis Property Fund Ltd	91224862	REBJ.J	ZAE000201687	2021	south africa
4	Crown Seal PCL	98936964	CSC.BK	TH0026010007	2023	thailand
5	Tata Investment Corporation Ltd	45838944	TINV.NS	INE672A01018	2023	india
6	Clarity Medical Group Holding Ltd	27380827	1406.HK	KYG2181S1084	2023	hong kong
7	Clarity Medical Group Holding Ltd	27380827	1406.HK	KYG2181S1084	2022	hong kong
8	Clarity Medical Group Holding Ltd	27380827	1406.HK	KYG2181S1084	2021	hong kong
9	Engineers India Ltd	392888202	ENGI.NS	INE510A01028	2023	india
10	ME Group International PLC	384549537	MEGPM.L	GB0008481250	2023	united kingdom
11	ME Group International PLC	384549537	MEGPM.L	GB0008481250	2022	united kingdom
12	ME Group International PLC	384549537	MEGPM.L	GB0008481250	2021	united kingdom
13	Works co uk PLC	361863773	WRKS.L	GB00BF5HBF20	2023	united kingdom
14	Works co uk PLC	361863773	WRKS.L	GB00BF5HBF20	2022	united kingdom

trbc_industry_group_code	trbc_industry_code_37	trbc_industry_code_38	scope1	scope2	scope3
601020	60102020	6010202010	7.20954e+02	111887.00	NA
601020	60102020	6010202010	NA	NA	NA
601020	60102020	6010202010	9.67930e+03	135468.07	NA
513020	51302010	5130201014	1.34550e+04	10817.00	44539.000
551020	55102020	5510202010	0.00000e+00	41.00	486014.000
561020	56102010	5610201011	4.70106e+07	4932140.00	NA
561020	56102010	5610201011	4.99289e+07	4931660.00	NA
561020	56102010	5610201011	NA	NA	NA
522010	52201020	5220102010	1.39000e+02	6802.00	NA
521020	52102010	5210201010	2.36000e+02	343.00	41368.000
521020	52102010	5210201010	3.35000e+02	402.00	27252.000
521020	52102010	5210201010	4.12150e+02	379.24	20919.500
534030	53403090	5340309010	1.99110e+02	2575.87	29.870

MÉTHODOLOGIE DE CALCUL TRADITIONNELLE

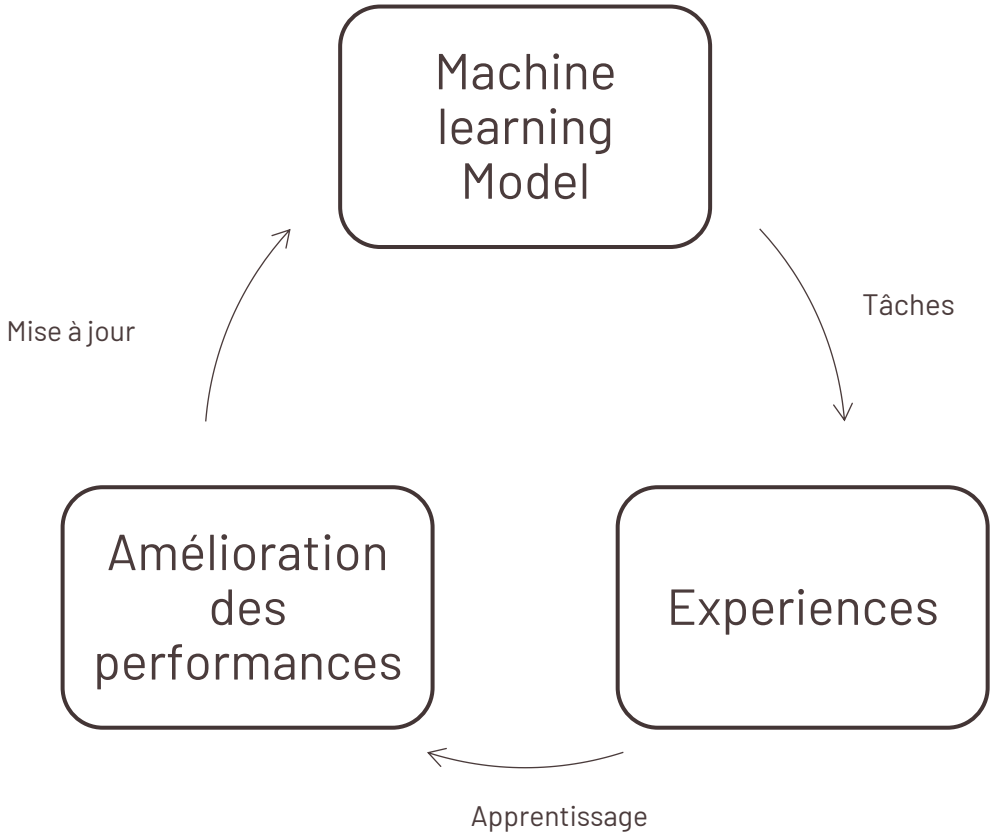
Pour chaque contrepartie:



LES FONDEMENTS DU PRINCIPE DE PRÉDICTION PAR LES RÉGRESSIONS

LA RÉGRESSION LINÉAIRE

- **L'objectif est de trouver les coefficients optimaux qui minimisent la différence entre la valeur prédite et la valeur réelle.**
- Collecte et préparation des données
- Division des données
- Entrainement du modèle de prédiction
- Évaluation du modèle : métrique de performance
- Ajustement et optimisation (ajout des variables, transformation en log utilisations de régressions etc.)
- Prédiction sur de nouvelles données



PRÉSENTATION DES MODÈLES

SPÉCIFICATION DES MODÈLES DE REGRESSION

Deux modèles ont été mis en œuvre pour l'année 2023. L'objectif de la prediction est de minimiser la distance entre la valeur des emissions réelles et la valeur des emissions prédites.

Taille de l'échantillon : après le traitement des données environ 7000 observations dans la base d'apprentissage et environ 700 observations dans la base de test.

Le modèle en niveau

- $émissions_{scope12} = \beta_0 + \beta_1 Revenu + \beta_2 Secteur + \beta_3 Region + \varepsilon_{it}$

Le modèle en log :

- $\log(émissions_{scope12}) = \beta_0 + \beta_1 \log(Revenu) + \beta_2 Secteur + \beta_3 Region + \varepsilon_{it}$

Le modèle avec des variables interaction (interaction1: Region*Secteur), interaction 2: region*Secteur*Revenu)

- $\log(émissions_{scope12}) = \beta_0 + \beta_1 \log(Revenu) + \beta_2 Secteur + \beta_3 Region + \beta_4 \log(Revenu) * Secteur + \beta_5 \log(Revenu) * Region + Secteur * Region + \varepsilon_{it}$

Ici les variables d'interaction représentent l'intèraction entre le revenu et les 42 secteurs pris séparément. Il en est de même pour les régions.

LES MODÈLES DE RÉGRESSIONS PÉNALISÉES

PRINCIPES → N < P

- Le Lasso ajoute une pénalité sur la somme des valeurs absolues des coefficients. Ce **qui peut à éliminer certaines variables pour simplifier le modèle à construire**

- Le Ridge Ajoute une pénalité à la somme des carrées des coefficients **empêchant les coefficients de devenir trop grand et aidant à gérer le problème de multi colinéarité**

- Le Elastic net: Combine les deux modèles pour obtenir les avantages du Lasso et du Ridge

Le choix du terme de pénalité se fait par le biais de la validation croisé. Part du principe de la limitation biais afin de trouver les coefficients les plus optimaux possibles en présence de multi colinéarité

Paramètre de la régression RIDGE

$$\min_{\beta_1,\cdots,\beta_p} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j z_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

λ (λ ≥ o) est un paramètre (coefficient de pénalité) qui permet de contrôler l'impact de la pénalité : à fixer

Fonction de pénalité

Paramètre de la régression LASSO

$$+ \lambda_1 \sum_{j=1}^{p} |\beta_j|$$

LES MODÈLES DE RÉGRESSIONS PÉNALISÉES

Avantages :

- Réduction du surapprentissage
- Le terme de pénalité permet de limiter la complexité du modèle en réduisant l'importance des coefficients
- Améliore la généralisation du modèle sur de nouvelles données

Inconvénients:

- Peuvent introduire des biais dans les étapes de la modélisation
- Elles modifient des coefficients estimés pour les rendre plus petits
- Le compromis bais-variance peut conduire à des performances sous optimales si le lambda est trop élevé

RÉSULTATS

MÉTRIQUE DE PERFORMANCE : ROOT MEAN SQUARE ERROR (RMSE)

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y})^2}$$

- Métrique qui quantifie l'écart type des erreurs de prédictions (erreurs résiduelles). Plus le RMSE est bas, plus le modèle est précis dans ces prédictions.
- Elle s'exprime dans la même unité que la variable de réponse (émissions scope 1 &2 2023).
- Aide à évaluer la capacité , à évaluer la capacité à générer sur de nouvelles données et éviter le surapprentissage.

MODÈLE EN NIVEAU

RMSE	Médiane	OLS	Ridge	Lasso	Elastic net
	> 5 millions	> 71 millions	> 5 millions	> 5 millions	> 5 millions

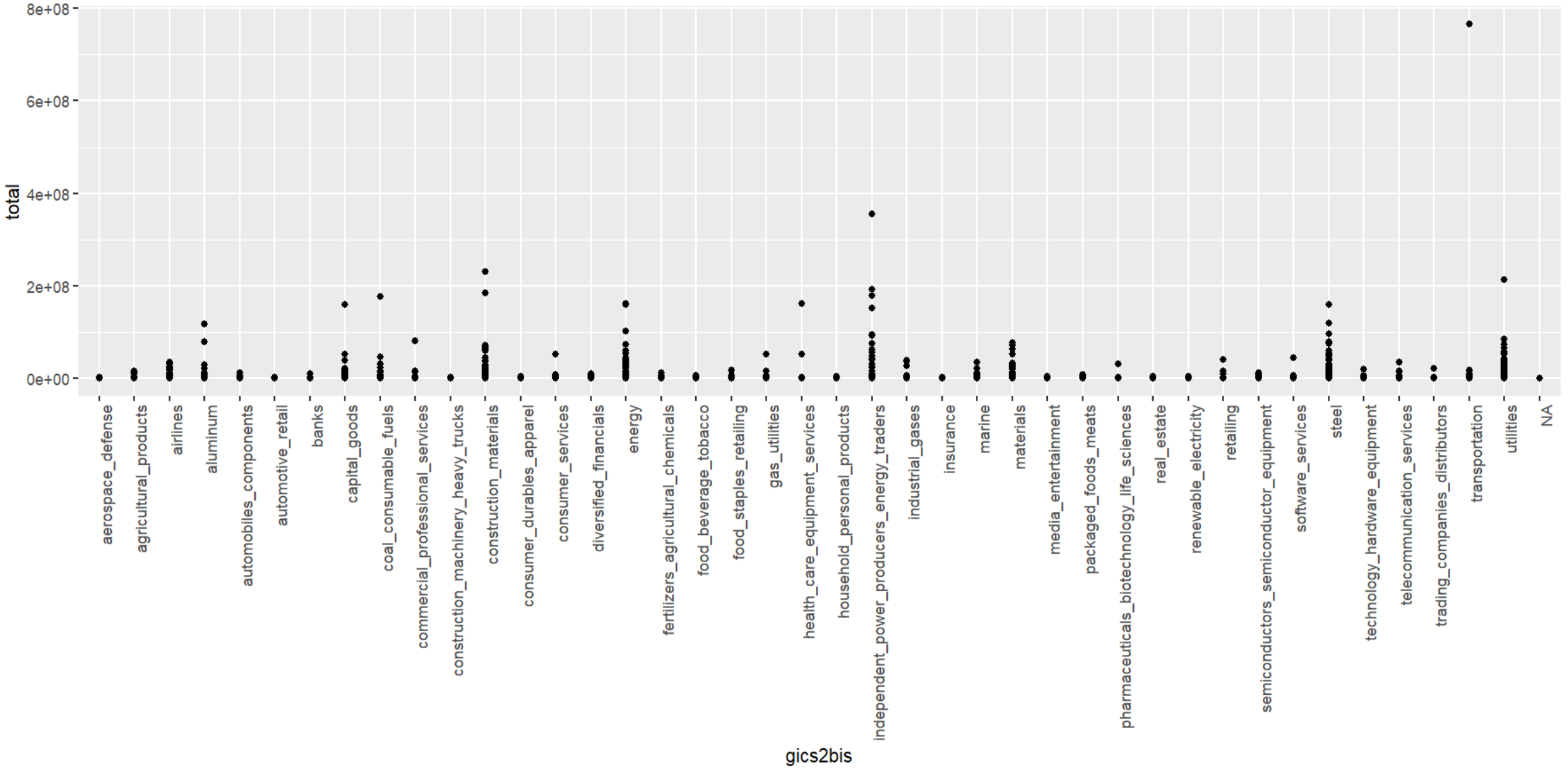
- **Conclusion**: le modèle à niveau présente une grande variabilité dans les données avec une présence des valeurs extrêmes. Ce qui conduit à surajustement dans les coefficients du modèles OLS

MODÈLE EN LOG

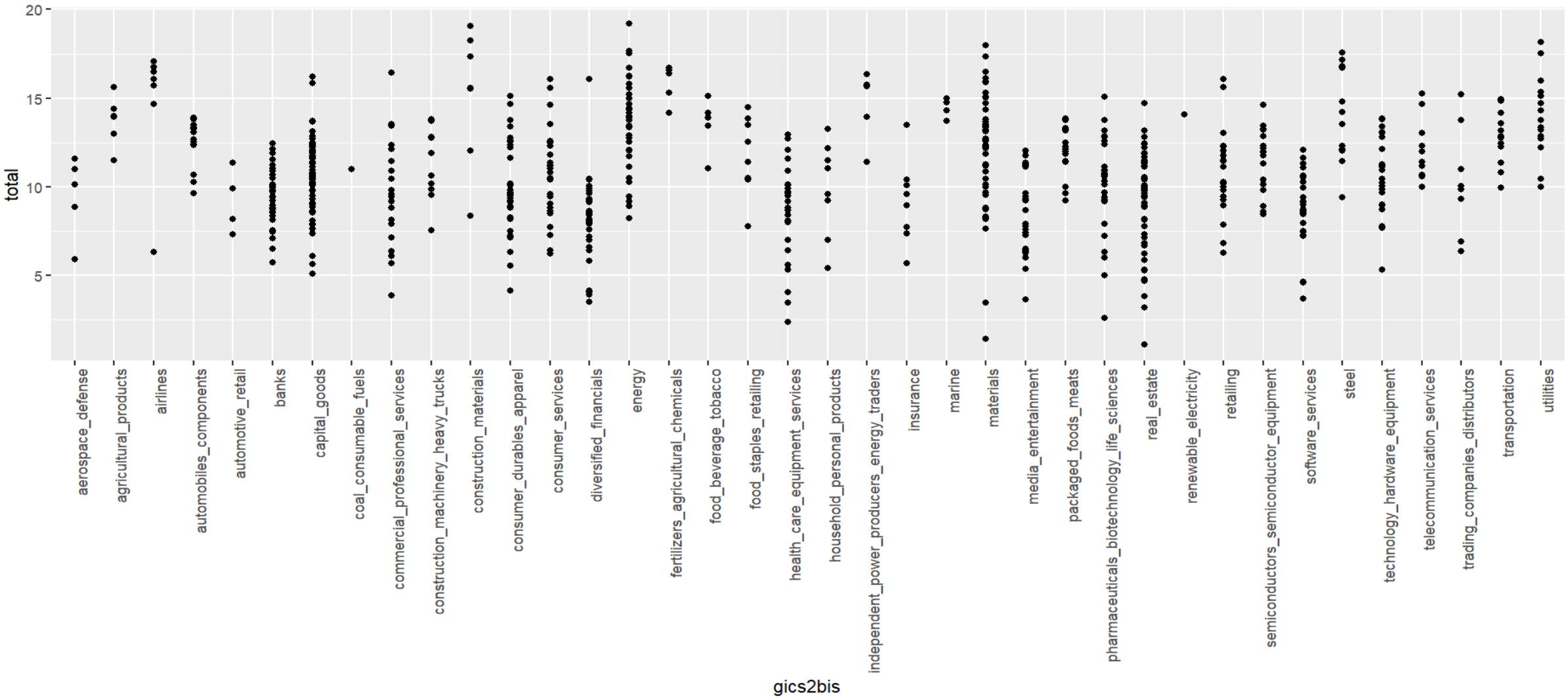
RMSE	Médiane	OLS	Ridge	Lasso	Elastic net
	1,88	1,83	1,83	1,83	1,83

- **Conclusion** : réduction de l'élasticité par les log et amélioration de la performance dans tous les modèles

ÉMISSIONS PAR SECTEUR SANS LOG



ÉMISSIONS PAR SECTEUR AVEC LOG



RÉSULTATS

MÉTRIQUE DE PERFORMANCE : ROOT MEAN SQUARE ERROR (RMSE)

MODÈLE EN LOG AVEC LES TERMES D'INTERACTIONS REGION* SECTEUR

RMSE	OLS	Ridge	Lasso	Elastic net
	1,83	1,81	1,82	1,82

MODÈLE EN LOG AVEC TOUTES LES INTERACTIONS

RMSE	OLS	Ridge	Lasso	Elastic net
	1,85	1,80	1,80	1,80

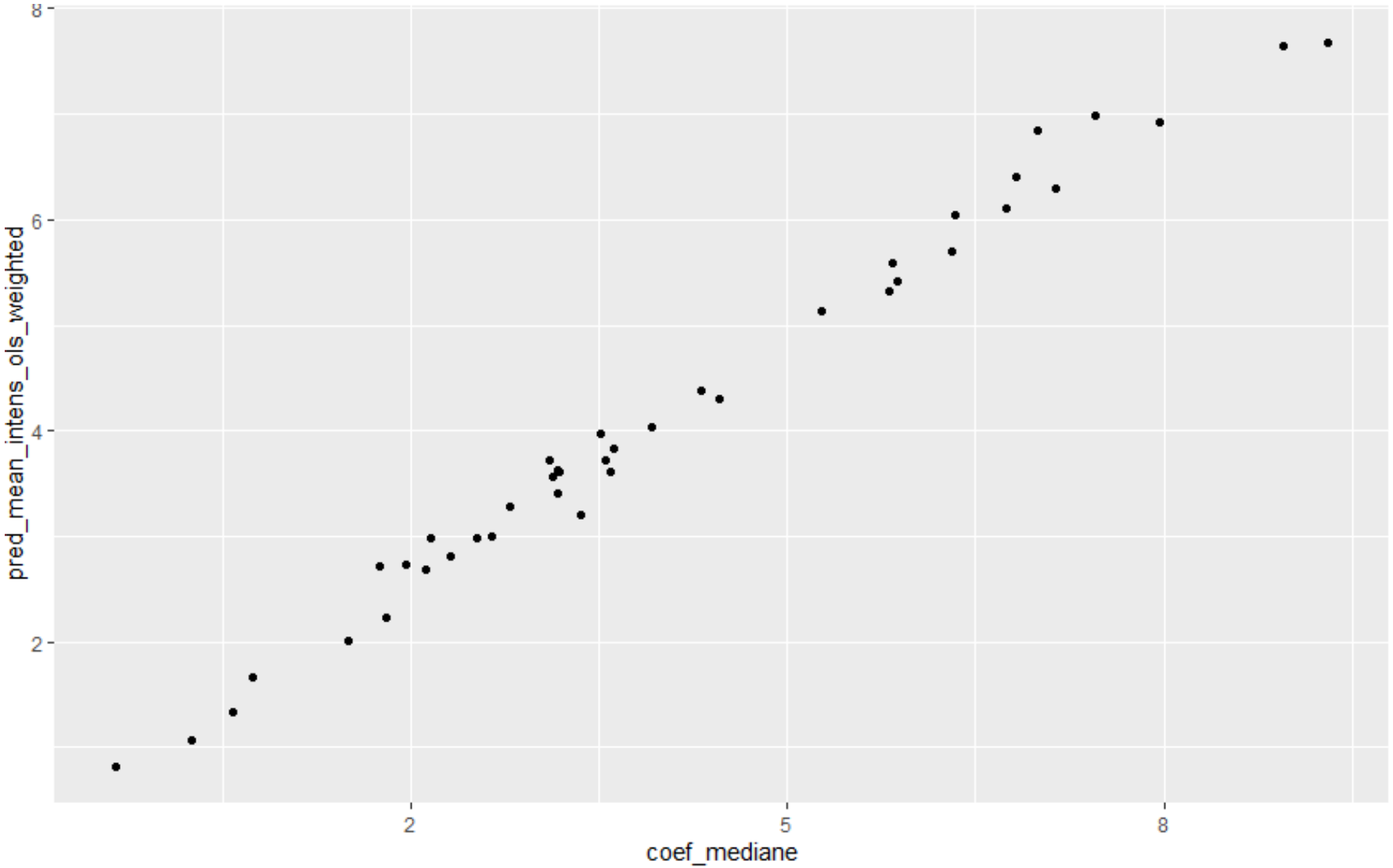
Conclusion : amélioration de la performance des modèles de régressions pénalisées. Plus il a des termes d'interaction, les modèles de régressions pénalisées permettent d'éviter le surapprentissage.

- Moyenne des émissions scope 1&2 2023 = 10.9
- Médiane des émissions scope 1 &2 2023 = 10.7

Conclusion : Pas de grande différence entre la moyenne et la médiane des émissions scope1&2 2023.

RÉSULTATS

- **COMPARAISON DES ÉMISSIONS SCOPE 1&2 PRÉDITES : MÉDIANE VS OLS**



Conclusion:

Similarité dans la prédiction des émissions scope 1 & 2 2023 par la médiane et par la régression linéaire

RÉSULTATS

Pays	Secteur	Entreprises	émissions12_2023	prediction_ols	prediction_ols_model_inter1	prediction_ols_model_inter2	Meilleure_prediction
france	retailing	Fnac Darty SA	9,99	10,87	10,87	10,87	OLS Inter1
france	consumer_services	Elior Group SA	11,21	11,43	11,43	11,21	OLS Inter2
france	consumer_durables_apparel	Kaufman & Broad SA	7,23	9,60	9,60	9,46	OLS Inter2
france	automobiles_components	Valeo SE	13,47	12,82	12,82	13,15	OLS Inter2
france	food_staples_retailing	Casino Guichard Perrachon SA	13,84	11,89	11,89	12,18	OLS Inter2
france	pharmaceuticals_biotechnology_life_sciences	Transgene SA	5,01	5,67	5,67	5,55	OLS Inter2
france	pharmaceuticals_biotechnology_life_sciences	Sanofi SA	13,19	13,21	13,21	13,09	OLS Inter1
france	commercial_professional_services	Seche Environnement SA	13,46	9,38	9,38	9,20	OLS Inter1
france	capital_goods	Exail Technologies	6,09	9,07	9,07	8,77	OLS Inter2
france	materials	Oeneo SA	10,01	10,81	10,81	11,09	OLS Inter1
france	materials	Eramet SA	15,08	12,85	12,85	13,17	OLS Inter2
france	real_estate	Nexity SA	8,95	10,71	10,71	10,91	OLS Inter1

CONCLUSION ET LIMITES

- Nombre de variable limitées (le chiffre d'affaires, le secteur, la région, émissions
- Problème de surapprentissage dans la régression pénalisée lorsqu'il a peu de variables explicatives
- Biais de représentativité des entreprises dans la base Reuters
- Le nombre de variable d'interaction améliore les performances des modèles mais trop de variables peuvent entrainer des bugs au niveau de l'ordinateur

EXTENSIONS SOUHAITÉES

- Imputation itérative MICE (Multiple Imputation by Chained Equation).
  - Imputation initiale : moyenne , médiane
  - Imputation itérative : Chaque variable contenant des NA sont considérés comme des variables dépendantes et les autres comme des variables explicatives.
  - Une fois une variable imputée, la procédure passe à la suivante et réitère jusqu'à ce que toutes les variables aient été traités.
- Créer des clusters des entreprises avec une imputation itérative par Cluster
  - Combiner MICE avec des algorithmes de clustering comme les K-means
  - L'objectif serait de créer des profils similaires d'entreprises par secteur dans reuters et le comparer avec les profils des entreprises par secteur contenu dans le portefeuille de BPIfrance
- Tester des nouvelles prédictions sur ces nouvelles méthodes

DIFFICULTÉS RENCONTRÉES

- ❑ Choix la spécification des modèles
  - Passer d'un modèle en niveau à un modèle en log
  - Prise en compte de la corrélation entre les variables.
  - Sélectionner les bonnes méthodes de nettoyage, de normalisation et de standardisation des données
- ❑ Choix de la bonne méthodologie à adopter
- ❑ Identification de la problématique
  - Formuler clairement la question de recherche ou le problème à résoudre.
  - Acquérir une bonne compréhension des données disponibles et de leurs caractéristiques pour pouvoir formuler une problématique pertinente.
- ❑ Difficultés techniques
  - Implémenter correctement les algorithmes de machine learning ou de régression
  - Mettre en place une validation croisée avec caret ou glmnet
  - Trouver le bon compromis spécification des modèles, les variables, l'implantation ML et évaluation peut nécessiter de nombreux essais, des blocages et des erreurs

MERCI